Are Slepian-Wolf Rates Necessary for Distributed Parameter Estimation?

Mostafa El Gamal and Lifeng Lai
Department of Electrical and Computer Engineering
Worcester Polytechnic Institute
{melgamal, llai}@wpi.edu

***Abstract*—We consider a distributed parameter estimation problem, in which multiple terminals send messages related to their local observations using limited rates to a fusion center who will obtain an estimate of a parameter related to observations of all terminals. It is well known that if the transmission rates are in the Slepian-Wolf region, the fusion center can fully recover all observations and hence can construct an estimator having the same performance as that of the centralized case. One natural question is whether Slepian-Wolf rates are necessary to achieve the same estimation performance as that of the centralized case. In this paper, we show that the answer to this question is negative. We establish our result by explicitly constructing an asymptotically minimum variance unbiased estimator (MVUE) that has the same performance as that of the optimal estimator in the centralized case while requiring information rates less than the conditions required in the Slepian-Wolf rate region.**

***Index Terms*—Distributed learning, MVUE, Slepian-Wolf rates, universal encoding/decoding scheme.**

I. Introduction

There are two main different setups for statistical learning: centralized learning and distributed learning. In the centralized learning, which has been studied extensively, all data is available at a centralized location. In the distributed learning, data is stored in multiple terminals. The distributed learning setup has attracted significant recent research interests as the data involved in learning is increasingly large in volume and might be stored in multiple terminals [1], [2], [3], [4]. For the distributed learning, each terminal either has a few observations about all variables, or has full knowledge about a subset of variables (all observations about a subset of variables). The first scenario is relatively easier since each terminal can still make its own local inference without even communicating with each other, while communication between terminals is essential for the second scenario. In this paper, we focus on the more challenging second scenario.

In particular, we consider a distributed parameter estimation problem. In the setup considered, there are two random variables (X, Y) with a joint probability mass function (PMF) $P_\theta(X, Y)$ parameterized by an unknown parameter θ. Two terminals A and B observe X^n and Y^n respectively and send messages related to their own local observations with limited rates to terminal C, which will then obtain an estimate of the unknown parameter. It is well known that if the transmission rates from the terminals are inside the Slepian-Wolf rate region [5], there exists a universal coding scheme [6] that enables terminal C to fully recover (X^n, Y^n). Hence, once the transmission rates are inside the Slepian-Wolf rate region, the performance of the best estimator for the distributed setup is the same as that of the best estimator for the centralized case.

One natural question is: are Slepian-Wolf rates *necessary* to achieve the same estimation performance as that of the centralized case? The answer to this question has significant implications in the distributed estimation. If the answer is yes, then to obtain the best estimate of the unknown parameter requires transmission rates to be so high that they are sufficient to fully recover the observations at the decoder, hence no rate reduction is possible. On the other hand, if the answer is no, then the observations can be compressed beyond the limits of source coding for full observation recovery. At a first glance, the answer to this question should be no as we are only interested in estimating a parameter related to the observations and are not interested in recovering the observations themselves. However, all existing related works indicate otherwise. For example, [7] addressed the same question and suggested that Slepian-Wolf rates might be necessary. In addition, the performance of the best known estimator by Han and Amari [8] does not match that of the centralized case when the information rates are outside of the Slepian-Wolf rate region. Furthermore, [9] showed that, under certain conditions, extracting even one bit of information from distributed sources is as hard as recovering full observations and hence requires the information rates to be in the Slepian-Wolf rate region.

In this paper, we show that the answer to this question is indeed *no*. We establish our result by explicitly constructing a distributed estimator that achieves the same performance as that of the optimal estimator for the centralized case while using information rates outside of the Slepian-Wolf region. In particular, we consider binary symmetric sources (i.e., both X^n and Y^n are binary sequences) parameterized by an unknown parameter θ. In our scheme, we first design a universal coding/decoding scheme that enables terminal C to compute $Z^n = X^n \oplus Y^n$, which can be achieved using rates outside of the Slepian-Wolf rate region, and then construct an estimator using Z^n. We show that our estimator is an asymptotically minimum variance unbiased estimator (MVUE) [10] and achieves the same variance index as that of the

The work of Mostafa El Gamal and Lifeng Lai was supported by the Qatar National Research Fund under Grant QNRF-6-1326-2-532.

best estimator in the centralized case. We further extend our scheme to a more general class of joint PMFs and show that our scheme can also achieve the same performance as that of the best estimator in the centralized case while using transmission rates less than the conditions required in the Slepian-Wolf rate region. The key idea of our scheme is, instead of fully recovering the source observations, we aim to recover sufficient statistics at terminal C using less information rates.

The rest of the paper is organized as follows. We introduce the problem formulation in Section II. In Section III, we establish our main results for the binary symmetric sources. We extend our work to a more general class of information sources in Section IV. We present the simulation results in Section V. Finally, we conclude the paper in Section VI.

II. Problem Formulation

Consider two information sources X and Y taking values from the discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. $(X^n, Y^n) = \{(X_i, Y_i)\}_{i=1}^n$ are n independently and identically distributed (i.i.d.) observations drawn according to the parametric joint PMF $P_\theta(X, Y)$ where $\theta \in \Theta$ is the unknown parameter. We consider a distributed setup in which X^n are observed at terminal A and Y^n are observed at terminal B. Using limited rates, these two terminals send messages related to their own local observations to a fusion center (terminal C), which will then obtain an estimate $\hat{\theta}$ of θ using these messages. The setup is illustrated in Fig. 1.

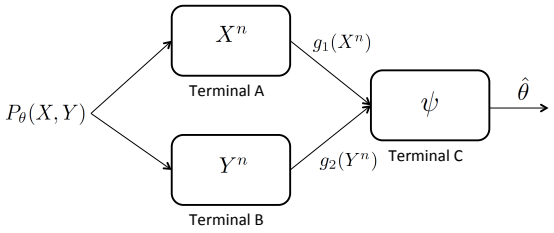

Fig. 1: System Model.

In particular, terminal A employs an encoding function $g_1 : X^n \to g_1(X^n)$, while terminal B employs an encoding function $g_2 : Y^n \to g_2(Y^n)$. The code rates are

$$R_X = \frac{\log ||g_1||}{n}, R_Y = \frac{\log ||g_2||}{n}, \tag{1}$$

where $||g_i||$ is the cardinality of the encoding function g_i.

From $g_1(X^n)$ and $g_2(Y^n)$, the decoder obtains an estimate $\hat{\theta}$ of the unknown parameter θ using estimator ψ:

$$\hat{\theta} = \psi(g_1(X^n), g_2(Y^n)). \tag{2}$$

To evaluate the quality of the estimator, we use the variance index that is defined as

$$V[\hat{\theta}] = \lim_{n\to\infty} n\text{Var}_\theta[\hat{\theta}] = \lim_{n\to\infty} n\mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]. \tag{3}$$

It is desirable to have an estimator that is asymptotically unbiased, i.e., $\mathbb{E}_\theta[\hat{\theta}] \to \theta$ as $n \to \infty$, and has a small variance index.

It is well-known that, if the coding rates satisfy (will be called Slepian-Wolf rates in the sequel)

$$R_X \geq H_\theta(X|Y), \tag{4}$$

$$R_Y \geq H_\theta(Y|X), \tag{5}$$

$$R_X + R_Y \geq H_\theta(X, Y), \tag{6}$$

there exists universal source coding schemes [6] (i.e., the coding scheme does not depends on the value of the unknown parameter θ) such that the decoder can reconstruct X^n and Y^n with a diminishing error probability. Here, $H_\theta(\cdot)$ and $H_\theta(\cdot|\cdot)$ denote the entropy and conditional entropy respectively[1]. Hence, if (4)-(6) are satisfied, we can obtain the same estimation performance as that of the centralized case.

The question we ask in this paper is: are Slepian-Wolf rates *necessary* to achieve the same estimation performance as that of the centralized case? [7] investigated the same question and suggested that Slepian-Wolf rates appear to be necessary for achieving the centralized estimation performance. In this paper, we show that Slepian-Wolf rates are *not* necessary. In particular, we show that there indeed exists a class of PMFs and the corresponding distributed estimators that require communication rates less than the Slepian-Wolf rates while still achieving the same performance as that of the best estimator for the centralized case.

Throughout the paper, we use an upper case letter Z to denote a random variable, and a lower case letter z to denote a realization of Z. For any sequence $z^n = (z(1), \cdots, z(n)) \in \mathcal{Z}^n$, the relative frequencies (empirical PMF) $\pi(a|z^n) \triangleq n(a|z^n)/n, \forall a \in \mathcal{Z}$ of the components of z^n is called the type of z^n. Here $n(a|z^n)$ is the total number of indices t at which $z(t) = a$.

III. Binary Symmetric Case

In this section, we consider the case of binary symmetric sources with $|\mathcal{X}| = |\mathcal{Y}| = 2$ and a joint PMF of (X, Y) as given in Table I, in which the unknown parameter $\theta \in \Theta = (0, 1)$.

X/Y	0	1
0	$\theta/2$	$(1-\theta)/2$
1	$(1-\theta)/2$	$\theta/2$

TABLE I: The joint PMF of binary symmetric sources.

We show that, to estimate θ for this class of PMFs, we can achieve the centralized estimation performance using rates that do not satisfy (4)-(6). We establish this result using two steps: 1) in the first step, we design a universal encoder at terminals A and B and universal decoder at terminal C to compute the modulo-two sum $Z^n = X^n \oplus Y^n$; 2) in the second step, we construct an estimator using Z^n.

[1]Throughout the paper, we use the subscript θ to emphasize the fact that value of the quantity of interest depends on the parameter θ.

A. Step 1: Comupting Z^n

Here, we discuss how to universally compute $Z^n = X^n \oplus Y^n$ at terminal C. Towards this goal, we will use the same linear code at both encoders and use a minimum entropy decoder at terminal C.

Since the encoders at terminals A and B are the same, we use the following simplified notation

$$\begin{aligned} f &= g_1 = g_2, \\ R &= R_X = R_Y. \end{aligned} \tag{7}$$

The following theorem shows that as long as $R \geq H_\theta(X|Y) = H_\theta(Y|X)$, the decoder can reconstruct Z^n with a diminishing error probability.

Theorem 1: If

$$R \geq H_\theta(X|Y) = H_\theta(Y|X), \tag{8}$$

there exist universal encoding/decoding functions to reconstruct $Z^n = X^n \oplus Y^n$ at terminal C with an exponentially decreasing error probability.

Proof: The proof follows a similar structure as the proofs in [11] and [6]. In particular, using the ideas in [6], we modify the proof of [11] to make it universal.

Random Code Generation: We use a linear code f with an encoding matrix A of size $n \times nR$ to map $\{0,1\}^n$ to $\{1, 2, ..., 2^{nR}\}$. Hence $||f|| = 2^{nR}$. We independently generate each entry of A using a uniform binary distribution, i.e., each entry of A is 0 or 1 with probability 0.5.

Encoding: The encoded messages of the realizations $x^n \in \{0,1\}^n$ and $y^n \in \{0,1\}^n$ are

$$\begin{aligned} f(x^n) &= x^n A, \\ f(y^n) &= y^n A. \end{aligned} \tag{9}$$

Decoding: The decoder first combines the messages into a single message as

$$f(x^n) \oplus f(y^n), \tag{10}$$

in which $\oplus$ denotes the element-wise xor.

It follows from the code linearity that

$$f(x^n) \oplus f(y^n) = f(x^n \oplus y^n) = f(z^n). \tag{11}$$

From $f(x^n \oplus y^n)$, terminal C uses a minimum entropy decoder to obtain $\hat{z}^n$. In particular, for each $\bar{z}^n$ such that $f(\bar{z}^n) = f(x^n \oplus y^n)$, the minimum entropy decoder first calculates the entropy of its type, then picks the one that has the least entropy to be the decoded sequence. In the following, to simplify the notation, we use $\bar{Z}^{(n)}$ and $Z^{(n)}$ to denote dummy random variables whose PMFs $P_{\bar{Z}^{(n)}}$ and $P_{Z^{(n)}}$ are the same as the types of $\bar{z}^n$ and z^n, respectively. The final decoded message is denoted as

$$\hat{z}^n = \phi(f(z^n)), \tag{12}$$

where ϕ denotes the minimum entropy decoding function.

Error Probability Analysis: A decoding error occurs if and only if there exists a sequence $\hat{z}^n \neq z^n$ such that

$$f(\hat{z}^n) = f(z^n) \;\text{ and }\; H(\hat{Z}^{(n)}) \leq H(Z^{(n)}). \tag{13}$$

The error probability, averaging over all possible codebooks, is

$$P_e^{(n)} = \sum_{z^n \in \{0,1\}^n} P_r(z^n) P_r(\hat{z}^n \neq z^n). \tag{14}$$

To analyze the probability of the decoding error, let $\tilde{z}^n \in \{0,1\}^n$ denote another sequence such that

$$\tilde{z}^n \neq z^n, \qquad f(\tilde{z}^n) = f(z^n). \tag{15}$$

Let $\tilde{Z}^{(n)}$ be a dummy random variable whose PMF $P_{\tilde{Z}^{(n)}}$ is the same as the type of $\tilde{z}^n$. Define $\mathcal{P}^{(n)}_{Z\tilde{Z}}$ as the set of all joint types between any two sequences z^n and $\tilde{z}^n$. For any given f (equivalently for a given encoding matrix A), define $N_f^n(Z\tilde{Z})$ as the number of sequences z^n such that there exists another sequence $\tilde{z}^n$ having the joint type $P_{Z^{(n)}\tilde{Z}^{(n)}} \in \mathcal{P}^{(n)}_{Z\tilde{Z}}$ and (15) holds.

Since each entry in A is uniformly distributed, then each element in $f(z^n)$ is uniformly distributed if z^n is a nonzero sequence. Therefore,

$$P_r(f(z^n) = 0) = (0.5)^{nR} = \frac{1}{||f||}, \tag{16}$$

in which the probability is computed over all codebooks. This implies that

$$P_r(f(\tilde{z}^n) = f(z^n)) = P_r(f(\tilde{z}^n - z^n) = 0) = \frac{1}{||f||}. \tag{17}$$

Define $T_{P_{Z^{(n)}\tilde{Z}^{(n)}}}$ as the set of all sequence pairs $(z^n, \tilde{z}^n)$ that have the joint type $P_{Z^{(n)}\tilde{Z}^{(n)}}$, $T_{P_{Z^{(n)}}}$ as the set of all sequences z^n that have the marginal type $P_{Z^{(n)}}$, and $T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)$ as the set of all sequences $\tilde{z}^n$ that have the joint type $P_{Z^{(n)}\tilde{Z}^{(n)}}$ with z^n. The sizes of the sets $T_{P_{Z^{(n)}}}$ and $T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)$ are bounded as [12]

$$\begin{aligned} |T_{P_{Z^{(n)}}}| &\leq 2^{nH(Z^{(n)})}, \\ |T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)| &\leq 2^{nH(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon}, \end{aligned} \tag{18}$$

where ϵ is an arbitrary small number. Notice that, for any given $P_{Z^{(n)}\tilde{Z}^{(n)}}$, $N_f^n(Z\tilde{Z})$ is a random variable (random over f) that can be expressed as

$$\begin{aligned} N_f^n(Z\tilde{Z}) &= \sum_{z^n \in T_{P_{Z^{(n)}}}} \mathbf{1}\big(\exists \tilde{z}^n \neq z^n : f(\tilde{z}^n) = f(z^n), \\ &\qquad \text{and } (z^n, \tilde{z}^n) \in T_{P_{Z^{(n)}\tilde{Z}^{(n)}}}\big) \\ &= \sum_{z^n \in T_{P_{Z^{(n)}}}} \mathbf{1}\big(\exists \tilde{z}^n \neq z^n : f(\tilde{z}^n) = f(z^n), \\ &\qquad \text{and } \tilde{z}^n \in T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)\big), \end{aligned} \tag{19}$$

where $\mathbf{1}(\cdot)$ is the indication function. The expectation of

$N_f^n(Z\tilde{Z})$ over all possible codebooks f is

$$\begin{aligned}
&\mathbb{E}[N_f^n(Z\tilde{Z})] \\
&= \sum_{z^n \in T_{P_{Z^{(n)}}}} \mathbb{E}\big[\mathbf{1}\big(\exists\ \tilde{z}^n \neq z^n\ :\ f(\tilde{z}^n) = f(z^n), \\
&\qquad\qquad \text{and } \tilde{z}^n \in T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)\big)\big] \\
&\leq \sum_{z^n \in T_{P_{Z^{(n)}}}} \sum_{\tilde{z}^n \in T_{P_{\tilde{Z}^{(n)}|Z^{(n)}}}(z^n)} P_r(f(\tilde{z}^n) = f(z^n)). \qquad (20)
\end{aligned}$$

(17), (18), and (20) imply that

$$\mathbb{E}[N_f^n(Z\tilde{Z})] \leq \frac{2^{n(H(Z^{(n)})+H(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon)}}{||f||}. \qquad (21)$$

Applying the Markov's inequality, we have

$$\begin{aligned}
&P_r\left(N_f^n(Z\tilde{Z}) \geq \frac{2^{n(H(Z^{(n)})+H(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon)}(|\mathcal{P}_{Z\tilde{Z}}^{(n)}|+\delta)}{||f||}\right) \\
&\quad \leq \frac{1}{|\mathcal{P}_{Z\tilde{Z}}^{(n)}|+\delta}, \qquad (22)
\end{aligned}$$

where $|\mathcal{P}_{Z\tilde{Z}}^{(n)}|$ is the total number of possible joint types and δ is an arbitrary small number. To simplify the notation, let

$$B^n(Z\tilde{Z}) \triangleq \frac{2^{n(H(Z^{(n)})+H(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon)}(|\mathcal{P}_{Z\tilde{Z}}^{(n)}|+\delta)}{||f||}. \qquad (23)$$

Considering all joint types $P_{Z^{(n)}\tilde{Z}^{(n)}}$ simultaneously, the union bound and (22) imply that

$$\begin{aligned}
&P_r\left(N_f^n(Z\tilde{Z}) \leq B^n(Z\tilde{Z}),\ \forall P_{Z^{(n)}\tilde{Z}^{(n)}} \in \mathcal{P}_{Z\tilde{Z}}^{(n)}\right) \\
&\geq\ 1 - \sum_{1}^{|\mathcal{P}_{Z\tilde{Z}}^{(n)}|} \frac{1}{|\mathcal{P}_{Z\tilde{Z}}^{(n)}|+\delta} \\
&>\ 0. \qquad (24)
\end{aligned}$$

Since the probability in (24) is positive, then there exists a codebook f^* that the following equation holds for all joint types $P_{Z\tilde{Z}}$ simultaneously

$$N_{f^*}^n(Z\tilde{Z}) \leq \frac{2^{n(H(Z^{(n)})+H(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon)}(|\mathcal{P}_{Z\tilde{Z}}^{(n)}|+\delta)}{||f^*||}. \qquad (25)$$

As $||f^*|| = 2^{nR}$ and $|\mathcal{P}_{Z\tilde{Z}}^{(n)}| \leq (n+1)^4$, we further have

$$\begin{aligned}
&N_{f^*}^n(Z\tilde{Z}) \qquad (26) \\
&\leq ((n+1)^4+\delta)\ \ 2^{n(H(Z^{(n)})+H(\tilde{Z}^{(n)}|Z^{(n)})+\epsilon-R)}.
\end{aligned}$$

In the following, we will focus on f^*.

Let $P_{e,f^*}^{(n)}(Z\tilde{Z})$ denote the portion of error probability associated with a fixed joint type $P_{Z^{(n)}\tilde{Z}^{(n)}}$

$$\begin{aligned}
&P_{e,f^*}^{(n)}(Z\tilde{Z}) \qquad (27) \\
&\triangleq \sum_{z^n \in T_{P_{Z^{(n)}}}} P_r(z^n)\mathbf{1}\big(\exists \tilde{z}^n \neq z^n : f^*(\tilde{z}^n) = f^*(z^n), \\
&\qquad\qquad \text{and } (z^n, \tilde{z}^n) \in T_{P_{Z^{(n)}\tilde{Z}^{(n)}}}\big).
\end{aligned}$$

The total decoding error probability $P_{e,f^*}^{(n)}$, when using f^*, can be expressed as

$$P_{e,f^*}^{(n)} = \sum_{P_{Z^{(n)}\tilde{Z}^{(n)}}} P_{e,f^*}^{(n)}(Z\tilde{Z}). \qquad (28)$$

Let $A_{\epsilon_1}^{(n)}$ denote the set of marginal types $P_{Z^{(n)}}$ such that $|P_{Z^{(n)}}(z=i) - P_\theta(z=i)| < \frac{\epsilon_1}{2}$ for $i \in \{0,1\}$, where ϵ_1 is an arbitrarily small number. Using the definition of $A_{\epsilon_1}^{(n)}$, (28) can be rewritten as

$$\begin{aligned}
P_{e,f^*}^{(n)} &= \sum_{P_{Z^{(n)}\tilde{Z}^{(n)}}, P_{Z^{(n)}} \in A_{\epsilon_1}^{(n)}} P_{e,f^*}^{(n)}(Z\tilde{Z}) \\
&\quad + \sum_{P_{Z^{(n)}\tilde{Z}^{(n)}}, P_{Z^{(n)}} \in \bar{A}_{\epsilon_1}^{(n)}} P_{e,f^*}^{(n)}(Z\tilde{Z}) \\
&\triangleq\ S_1 + S_2, \qquad (29)
\end{aligned}$$

where $\bar{A}_{\epsilon_1}^{(n)}$ denotes the complimentary set of $A_{\epsilon_1}^{(n)}$. For S_2, we have that

$$P_{e,f^*}^{(n)}(Z\tilde{Z}) \leq 2^{-n(D(P_{Z^{(n)}}||P_\theta(Z)))}, \qquad (30)$$

where $D(P_{Z^{(n)}}||P_\theta(Z))$ is the KL divergence between the marginal type $P_{Z^{(n)}}$ and the true PMF $P_\theta(Z)$ of $Z = X \oplus Y$. Using Pinsker's inequality, for $P_{Z^{(n)}} \in \bar{A}_{\epsilon_1}^{(n)}$, we have

$$D(P_{Z^{(n)}}||P_\theta(Z)) \geq 2\epsilon_1^2. \qquad (31)$$

Therefore,

$$\begin{aligned}
S_2 &\leq \sum_{P_{Z^{(n)}\tilde{Z}^{(n)}}} 2^{-2n\epsilon_1^2} \\
&\leq\ (n+1)^4\ \ 2^{-2n\epsilon_1^2}. \qquad (32)
\end{aligned}$$

(32) implies that $S_2 \to 0$ exponentially as $n \to \infty$.

For S_1, we have that

$$P_{e,f^*}^{(n)}(Z\tilde{Z}) \leq N_{f^*}^n(Z\tilde{Z})\ \ 2^{-n(H(Z^{(n)})+D(P_{Z^{(n)}}||P_\theta(Z)))}. \qquad (33)$$

Using (26), we further have

$$\begin{aligned}
&P_{e,f^*}^{(n)}(Z\tilde{Z}) \leq \qquad (34) \\
&((n+1)^4+\delta)\ 2^{-n\left(D(P_{Z^{(n)}}||P_\theta(Z))+R-H(\tilde{Z}^{(n)}|Z^{(n)})-\epsilon\right)}.
\end{aligned}$$

As we use the minimum entropy decoder, we have $H(\tilde{Z}^{(n)}) \leq H(Z^{(n)})$, which implies $H(\tilde{Z}^{(n)}|Z^{(n)}) \leq H(\tilde{Z}^{(n)}) \leq H(Z^{(n)})$. Therefore,

$$\begin{aligned}
&P_{e,f^*}^{(n)}(Z\tilde{Z}) \qquad (35) \\
&\leq ((n+1)^4+\delta)\ 2^{-n\left(D(P_{Z^{(n)}}||P_\theta(Z))+R-H(Z^{(n)})-\epsilon\right)}.
\end{aligned}$$

Since $P_{Z^{(n)}} \in A_{\epsilon_1}^{(n)}$, it is easy to check that

$$|H(Z^{(n)}) - H_\theta(Z)| \leq D(P_{Z^{(n)}}||P_\theta(Z)) + \epsilon_2. \qquad (36)$$

Here

$$\epsilon_2 = -\frac{\epsilon_1}{2}\sum_i \log P_\theta(z=i), \qquad (37)$$

which can be made arbitrarily small as $\epsilon_1 \downarrow 0$ for $\theta \in (0,1)$.

Therefore,

$$P_{e,f^*}^{(n)}(Z\tilde{Z}) \leq ((n+1)^4+\delta)\ 2^{-n\left(R-H_\theta(Z)-\epsilon_3\right)}, \quad (38)$$

in which $\epsilon_3 = \epsilon + \epsilon_2$.

This implies that $S_1 \to 0$ exponentially as $n \to \infty$ if

$$R \geq H_\theta(Z). \quad (39)$$

Therefore, (39) is sufficient to guarantee that $P_{e,f^*}^{(n)} \to 0$ exponentially as $n \to \infty$. It is easy to check that $H_\theta(Z) = H_\theta(X|Y) = H_\theta(Y|X)$. The proof is complete. ∎

Theorem 1 implies that the required rates to decode $Z^n = X^n \oplus Y^n$ with a small error probability is

$$R_X \geq H_\theta(X|Y), \quad (40)$$

$$R_Y \geq H_\theta(Y|X). \quad (41)$$

This rate region is larger than the Slepian-Wolf region in (4)-(6), as the condition $R_X + R_Y \geq H_\theta(X,Y)$ is not necessary anymore.

B. Step 2: Estimation

After obtaining $\hat{Z}^n$, which is equal to Z^n with a probability converging to 1 exponentially, we then design an asymptotically MVUE of θ. Our estimator is

$$\hat{\theta} = \frac{n(0|\hat{Z}^n)}{n}, \quad (42)$$

in which the notation $n(\cdot|\cdot)$ is defined in Section II.

Theorem 2: If the conditions in Theorem 1 are satisfied, the estimator in (42) is an asymptotically MVUE and achieves the optimal variance index as that of the centralized case.

Proof:

Consider the centralized case in which X^n and Y^n are both known perfectly. Let $\left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n}, \frac{n_4}{n}\right)$ denote the joint type of the sequences x^n and y^n, where (n_1, n_2, n_3, n_4) are the frequencies of occurrence of the pairs $\{(0,0),(1,1),(0,1),(1,0)\}$, respectively. The joint PMF of (x^n, y^n) is

$$P_\theta(x^n, y^n) = \left(\frac{\theta}{2}\right)^{(n_1+n_2)} \left(\frac{1-\theta}{2}\right)^{(n_3+n_4)} \quad (43)$$

Consider the centralized estimator

$$\hat{\theta}_c = \frac{(n_1+n_2)}{n}. \quad (44)$$

This estimator is unbiased since

$$\mathbb{E}_\theta[\hat{\theta}_c] = \theta. \quad (45)$$

The variance of the estimator is calculated as

$$\begin{aligned} \text{Var}_\theta[\hat{\theta}_c] &= \frac{1}{n^2}\mathbb{E}_\theta[(n_1+n_2)^2] - \theta^2 \\ &= \frac{\theta(1-\theta)}{n}. \end{aligned} \quad (46)$$

The variance index is given by

$$V[\hat{\theta}_c] = \lim_{n\to\infty} n\text{Var}_\theta[\hat{\theta}_c] = \theta(1-\theta). \quad (47)$$

The Cramer-Rao lower bound (CRLB) of the centralized case

$$\begin{aligned} \text{CRLB} &= -1/\mathbb{E}_\theta\left[\frac{\partial^2 \ln[P_\theta(x^n, y^n)]}{\partial^2\theta}\right] \quad (48) \\ &= \frac{\theta(1-\theta)}{n} = \text{Var}_\theta[\hat{\theta}_c]. \quad (49) \end{aligned}$$

This implies that $\hat{\theta}_c$ is an MVUE for the centralized case.

Now, come back to our decentralized case. For our estimator

$$\hat{\theta} = \frac{n(0|\hat{Z}^n)}{n}, \quad (50)$$

we have that

$$P_r(n(0|\hat{Z}^n) = n_1+n_2) \geq 1 - P_{e,f^*}^{(n)}, \quad (51)$$

in which $P_{e,f^*}^{(n)}$ is shown to converge to zero exponentially fast in Section III-A.

Therefore

$$\begin{aligned} \mathbb{E}_\theta[\hat{\theta}] = P_r(n(0|\hat{Z}^n) = n_1+n_2)\mathbb{E}_\theta[\hat{\theta}_c] \\ +(1-P_r(n(0|\hat{Z}^n) = n_1+n_2))K_1, \end{aligned} \quad (52)$$

where $K_1 \in [0,1]$ is a constant. As $n \to \infty$, $P_{e,f^*}^{(n)} \to 0$ and hence $P_r(n(0|\hat{Z}^n) = n_1+n_2) \to 1$. Therefore,

$$\lim_{n\to\infty} \mathbb{E}_\theta[\hat{\theta}] = \mathbb{E}_\theta[\hat{\theta}_c] = \theta. \quad (53)$$

This shows that our estimator is asymptotically unbiased. Similarly,

$$\begin{aligned} V[\hat{\theta}] &= \lim_{n\to\infty} n\text{Var}_\theta[\hat{\theta}] \\ &= \lim_{n\to\infty} (nP_r(n(0|\hat{Z}^n) = n_1+n_2)(\mathbb{E}_\theta[\hat{\theta}_c^2] - (\mathbb{E}_\theta[\hat{\theta}_c])^2) \\ &\quad +n(1-P_r(n(0|\hat{Z}^n) = n_1+n_2))(K_2 - K_1^2)), \end{aligned}$$

where $K_2 \in [0,1]$ is a constant. As $n \to \infty$, $P_{e,f^*}^{(n)} \to 0$ exponentially. Therefore,

$$V[\hat{\theta}] = \theta(1-\theta) = V[\hat{\theta}_c]. \quad (54)$$

This proves that our estimator is asymptotically unbiased and achieves the same minimum variance that can be achieved even in the centralized case. Hence, our estimator is optimal. ∎

Combining Theorems 1 and 2, we conclude that, in the distributed parameter estimation, the Slepian-Wolf rates are not necessary to achieve the same optimal estimation performance as that of the centralized case. Fig. 2 illustrates the comparison between the Slepian-Wolf rate region and the rate pair used in our estimator.

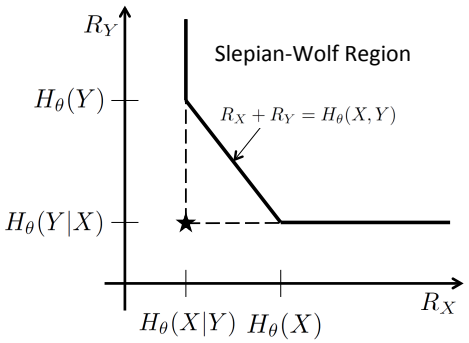

Fig. 2: ★: the rate pair required in our estimator, which is outside of the Slepian-Wolf rate region.

IV. Extension

In this section, we extend our results obtained in Section III to a more general class of joint PMFs. Let $\mathcal{X} = \mathcal{Y} = \{0, 1, ..., M-1\}$ and the class of PMFs be

$$P_\theta(X=i, Y=j) = \begin{cases} \frac{\theta}{M} & \text{, if } (i+j) \neq M-1 \\ \frac{1-\theta(M-1)}{M} & \text{, otherwise,} \end{cases} \tag{55}$$

where $\theta \in \Theta = (0, \frac{1}{(M-1)})$. Notice that each information source has a uniform marginal PMF and setting $M = 2$ recovers the binary case.

Similar to the binary case, we first use a linear code and minimum entropy decoder to reconstruct $Z^n = (X^n + Y^n) \mod M$ at the decoder and then design an estimator from Z^n. In this section, we use $\mod M$ to denote element-wise mod operation,

In particular, we use a linear code f that maps $\{0, 1, ..., M-1\}^n$ to $\{0, 1, ..., M-1\}^k$. The encoded messages of the realizations $x^n \in \{0, 1, ..., M-1\}^n$ and $y^n \in \{0, 1, ..., M-1\}^n$ are

$$\begin{aligned} f(x^n) &= x^n A, \\ f(y^n) &= y^n A, \end{aligned} \tag{56}$$

in which the code matrix A has n rows and k columns with each entry taking values from $\{0, 1, ..., M-1\}$. The coding rate is

$$R = \frac{k}{n} \log M. \tag{57}$$

The decoder first combines the encoded messages into a single message as

$$f(x^n) + f(y^n) \mod M. \tag{58}$$

The final decoded message is given by

$$\hat{z}^n = \phi(f(z^n)), \tag{59}$$

where ϕ the the minimum entropy decoding function. Following the same error probability analysis for the binary case, we can show that there exists a codebook f^* (and hence a particular encoding matrix A) that achieves a probability of decoding error $P_{e,f^*}^{(n)} \to 0$ exponentially as $n \to \infty$ if

$$R \geq H_\theta(Z) = H_\theta(X|Y) = H_\theta(Y|X). \tag{60}$$

Therefore, as long as

$$R_X \geq H_\theta(X|Y), \tag{61}$$

$$R_Y \geq H_\theta(Y|X), \tag{62}$$

we can reconstruct $Z^n = X^n + Y^n \mod M$ at the decoder with an exponentially diminishing error probability.

After obtaining $\hat{Z}^n$, which is equal to Z^n with a probability converging to 1 exponentially, our estimator is

$$\hat{\theta} = \frac{n - n(M-1|\hat{Z}^n)}{n(M-1)}. \tag{63}$$

Following similar steps as those in the binary case, we can show that, if (61)-(62) are satisfied, the estimator in (63) is asymptotically unbiased and achieves a variance index

$$V[\hat{\theta}] = \frac{\theta[1-\theta(M-1)]}{M-1}. \tag{64}$$

We can further show that (64) is the best variance index that can be achieved even in the centralized case. This implies that our scheme achieves the centralized performance using rates outside the Slepian-Wolf region.

V. Simulation Results

In this section, we compare our estimator to the best known estimator by Han and Amari [8]. In the simulation, we fix the unknown parameter θ and change the encoding rates R_X and R_Y such that

$$R_X = R_Y = R \geq H_\theta(Z). \tag{65}$$

We conduct the comparison for $M = 2$ and $M = 4$ respectively.

For $M = 2$, the variance index of our estimator is (54), while the variance index of the estimator by Han and Amari is calculated in example 3 of [8]

$$(\text{Var}_\theta[\hat{\theta}])_{HA} \simeq \frac{1}{16a^2b^2}\left\{\frac{1}{4} - \left(\theta - \frac{1}{2}\right)^2 [1 - (1-4a^2)(1-4b^2)]\right\}, \tag{66}$$

where a and b are functions of R_X and R_Y, whose expressions are given in (14.12) and (14.13) of [8], respectively.

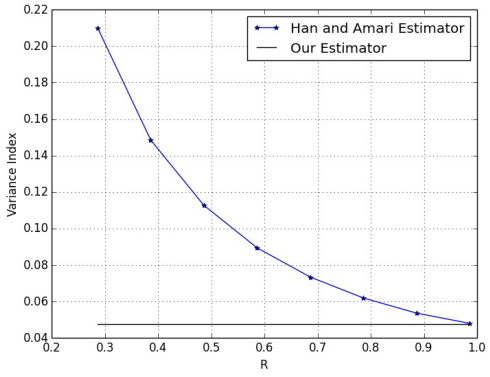

Fig. 3: Performance Comparison: $\theta = 0.05$, $M = 2$

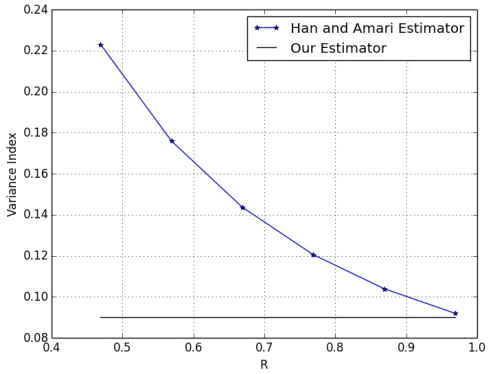

Fig. 4: Performance Comparison: $\theta = 0.9$, $M = 2$

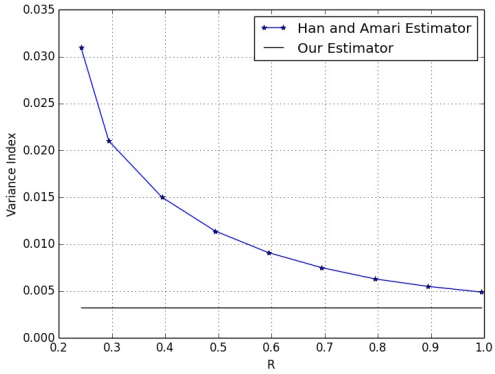

Fig. 5: Performance Comparison: $\theta = 0.01$, $M = 4$

Fig. 3 and Fig. 4 show the performance gain, in terms of the variance index, of our estimator over Han and Amari's estimator for binary symmetric sources ($M = 2$) at two different values of the unknown parameter, $\theta = 0.05$ and $\theta = 0.9$, respectively. The performance difference is more noticeable at low rates. For $\theta = 0.05$, the Slepian-Wolf sum rate is $R_X + R_Y = 1.29$ bits, while our estimator requires a sum rate of $R_X + R_Y = 2R = 0.57$ bits. For $\theta = 0.9$, the Slepian-Wolf sum rate is 1.47 bits, while our estimator requires a sum rate of 0.94 bits. Furthermore, for Han and Amari's estimator to achieve the centralized performance, the required sum-rate is 2 bits for both cases, which is not only much larger than the sum rate required in our estimator but also much larger than the sum-rate required by conditions specified in the Slepian-Wolf rate region.

For $M = 4$, the variance index of our estimator is given in (64). The performance of Han and Amari's estimator relies on the choice of the test channels. The authors did not specify an optimal choice of the test channels in order to extend example 3 in [8] to the case of $M = 4$. We find the following mapping to be a natural extension:

$$Q = \begin{cases} 0, \text{ if } & X \in \{0,1\} \\ 1, \text{ if } & X \in \{2,3\}, \end{cases} \qquad T = \begin{cases} 0, \text{ if } & Y \in \{0,1\} \\ 1, \text{ if } & Y \in \{2,3\}. \end{cases} \tag{67}$$

Notice that (Q, T) are distributed according to a binary symmetric PMF with an unknown parameter $\alpha = 2\theta$. Using an estimator $\hat{\theta} = \frac{\hat{\alpha}}{2}$ leads to the following expression for the variance index:

$$(\mathrm{Var}_\theta[\hat{\theta}])_{HA} \simeq \tag{68}$$
$$\frac{1}{64a^2b^2}\left\{\frac{1}{4} - \left(2\theta - \frac{1}{2}\right)^2 [1 - (1-4a^2)(1-4b^2)]\right\}.$$

Fig. 5 compares the variance indices achieved using our estimator and Han and Amari's estimator for $M = 4$ and $\theta = 0.01$. It is clear that our estimator outperforms that of Han and Amari's estimator. Furthermore, the performance difference is more noticeable at low rates. The Slepian-Wolf sum rate is 2.24 bits, while our estimator requires a sum rate of 0.48 bits.

VI. Conclusion

In this paper, we have answered the question: Are Slepian-Wolf rates necessary to achieve the same estimation performance as that of the centralized case? We have showed that the answer to this question is negative by constructing an asymptotically MVUE for binary symmetric sources using rates less than the conditions required in the Slepian-Wolf rate region. We have also extended our work to a general class of information sources by modifying the encoding/decoding scheme and the estimation algorithm. We have further compared our results to the best known estimator by Han and Amari to show the superiority of our estimator.

References

[1] M. Raginsky, "Learning from compressed observations," in *Proc. IEEE Inform. Theory Workshop*, (Tahoe City, CA), pp. 420–425, Sep. 2007.

[2] A. Xu and M. Raginsky, "Converses for distributed estimation via strong data processing inequalities," in *Proc. IEEE Intl. Symposium on Inform. Theory*, (Hong Kong, China), Jun. 2015.

[3] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems*, (Stateline, NV), pp. 2328–2336, Dec. 2013.

[4] O. Shamir and N. Srebro, "Distributed stochastic optimization and learning," in *Proc. Allerton Conf. on Communication, Control, and Computing*, (Monticello, IL), pp. 850–857, Oct. 2014.

[5] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, Nov. 1973.

[6] I. Csiszár, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Trans. Inform. Theory*, vol. 28, pp. 585–592, Jul. 1982.

[7] A. Zia, J. P. Reilly, and S. Shirani, "Distributed estimation; three theorems," in *Proc. IEEE Inform. Theory Workshop*, (Tahoe City, CA), pp. 517–522, Sep. 2007.

[8] T. S. Han and S.-I. Amari, "Parameter estimation with multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1802–1833, Nov. 1995.

[9] R. Ahlswede and I. Csiszár, "To get a bit of information may be as hard as to get full information," *IEEE Trans. Inform. Theory*, vol. 27, pp. 398–408, Jul. 1981.

[10] H. V. Poor, *An introduction to signal detection and estimation*. New York: Springer Science & Business Media, 2013.

[11] J. Körner and K. Marton, "How to encode the modulo-two sum of binary sources," *IEEE Trans. Inform. Theory*, vol. 25, pp. 219–221, Mar. 1979.

[12] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2505–2523, Oct. 1998.